BERT: PRE-TRAINING OF DEEP BIDIRECTIONAL TRANSFORMER FOR LANGUAGE UNDERSTANDING

GOOGLE AI LANGUAGE

PRESENTED BY: 艾查妮

OUTLINE

- Pre-requisite Knowledge
- Introduction to BERT
- BERT Architecture
- Transformer Encoder
- Pre-Training Procedure
- Experiments
- Results
- Conclusion

INTRODUCTION TO NLP

WHAT IS NATURAL LANGUAGE PROCESSING?

- A field of computer science, artificial intelligence and computational linguistics to perform useful tasks involving human languages
  - Human to Machine communication
  - Improving human-human communication
  - Extracting information from texts.
- Highly ambiguous
- Sentence *I made her duck* may have different meanings
  - I cooked waterfowl for her
  - I cooked waterfowl belong to her
  - I created the (plaster) duck she owns.
  - I caused her to quickly lower her head or body.

WHY IS NLP HARD?

Word segmentation

Part of Speech (POS)

Lexical Analysis

It is the vocabulary that includes words and expressions

It divides a text into paragraphs, sentences and words

Syntactic Analysis

It involves analysis of words in a sentence to depict the grammatical structure of the sentence

"The girl the go to the school"

Semantic Analysis

It abstracts the dictionary meaning or exact the meaning from the context

"colorless blue idea"

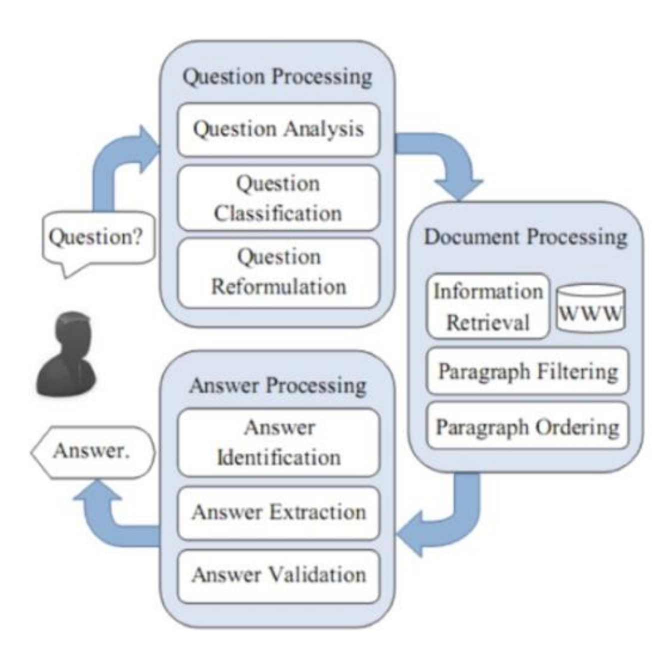
Discourse integration

Sense the context based on the previous sentence

"She wanted it"

NLP APPLICATIONS

Question Answering



Text Summarization

Distilling the most important information from text

Categories

- Single/Multi-document Summarization
- Extractive/Abstractive Summarization
- Query-focused text Summarization

Machine Translation

The process of translating one language to the other one.

Google Translation

The future looks surprisingly bright due to his hard work.

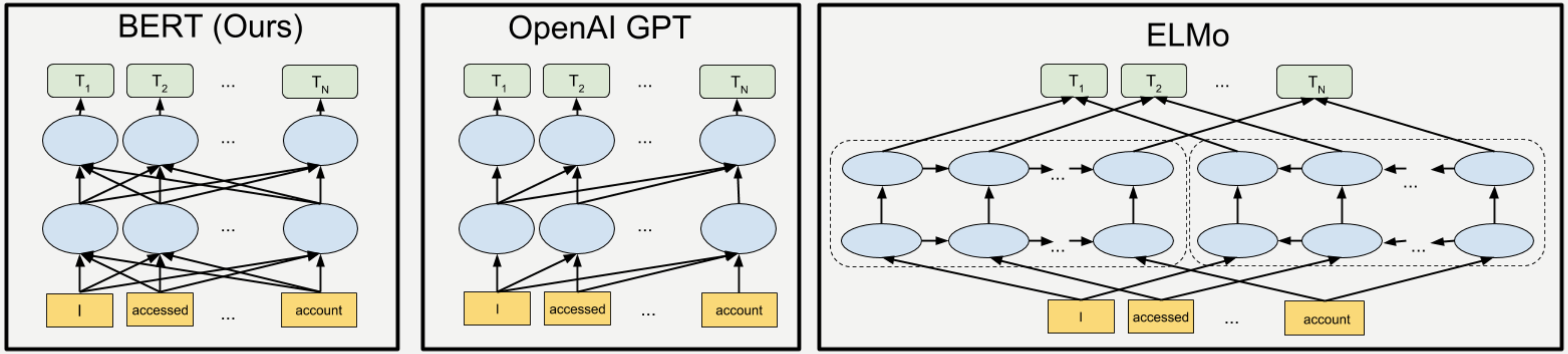
由於他的辛勤工作，未來看起來非常光明

LIMITATIONS OF CURRENT TECHNIQUES

- There are two existing strategies for applying pre-training language representation:

    → Feature based (ELMO)

    → Fine tuning (Open AI GPT, Generative Pre-trained Transformer)



BERT: BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMER

BERT

Bidirectional

It can read from both directions left as well as right to gain better understanding of the text

Encoder

This architecture is already well known as Encoder-Decoder for NLP tasks e.g. Seq2Seq and Machine Translation

Representation

Encoder-Decoder architecture is represented as Transformer

Transformer

key component is Multi-head attention block. It is combination of attention + normalization +masked attention in decoder phase

BERT: BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMER

- Main Ideas
  - Propose a new pre-training objective so that a deep bidirectional transformer can be trained.
    - The "masked language model" (MLM): the objective is to predict the original word of a masked word based only on its context.
    - "Next sentence prediction"
- Merits of BERT
    - Fine-tune BERT model for specific tasks to achieve state-of-the-art performance.
    - BERT advances the state-of-the-art for 11 NLP tasks

BERT ARCHITECTURE

- BERT's model architecture is a multi-layer bidirectional transformer encoder
  - (Vaswani et al., 2017) "Attention is all you need"
- Two models with different sizes were investigates
  - $BERT_{base}$: L=12, H=768, A=12, Total Parameters = 110M
  - $BERT_{large}$: L=24, H=1024, A=16, Total Parameters=340M

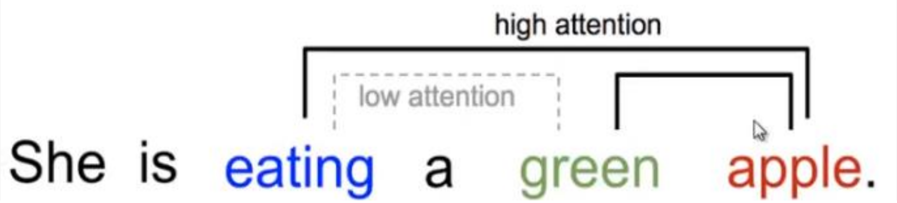
    L: number of layers (Transformer blocks),

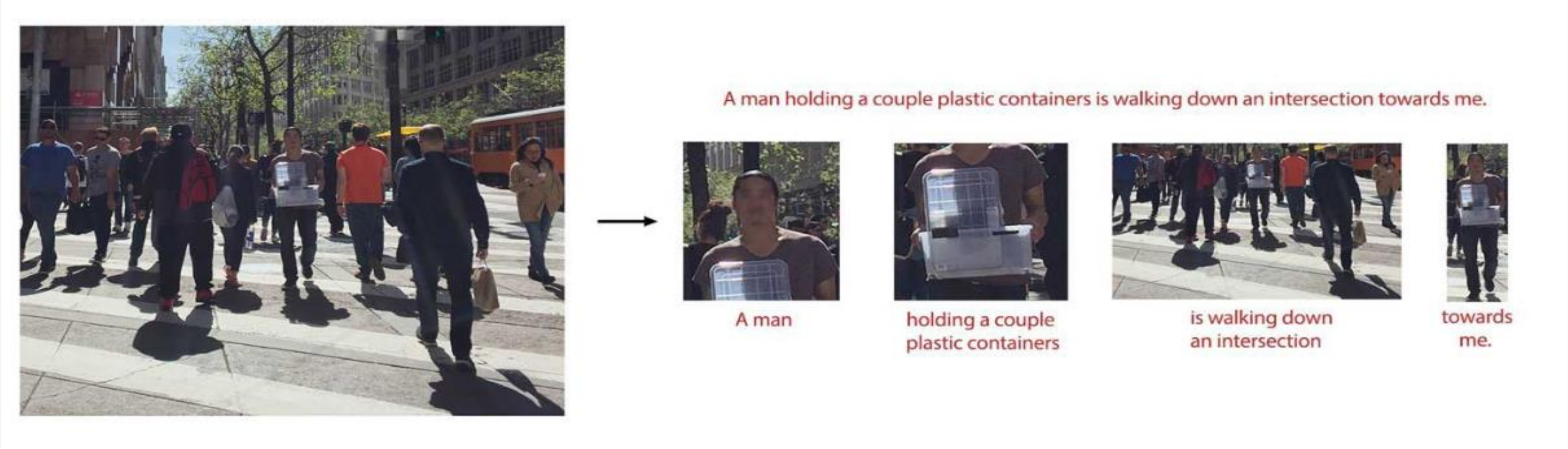
    H: hidden size

    A: number of self-attention heads

CONCEPT OF ATTENTION

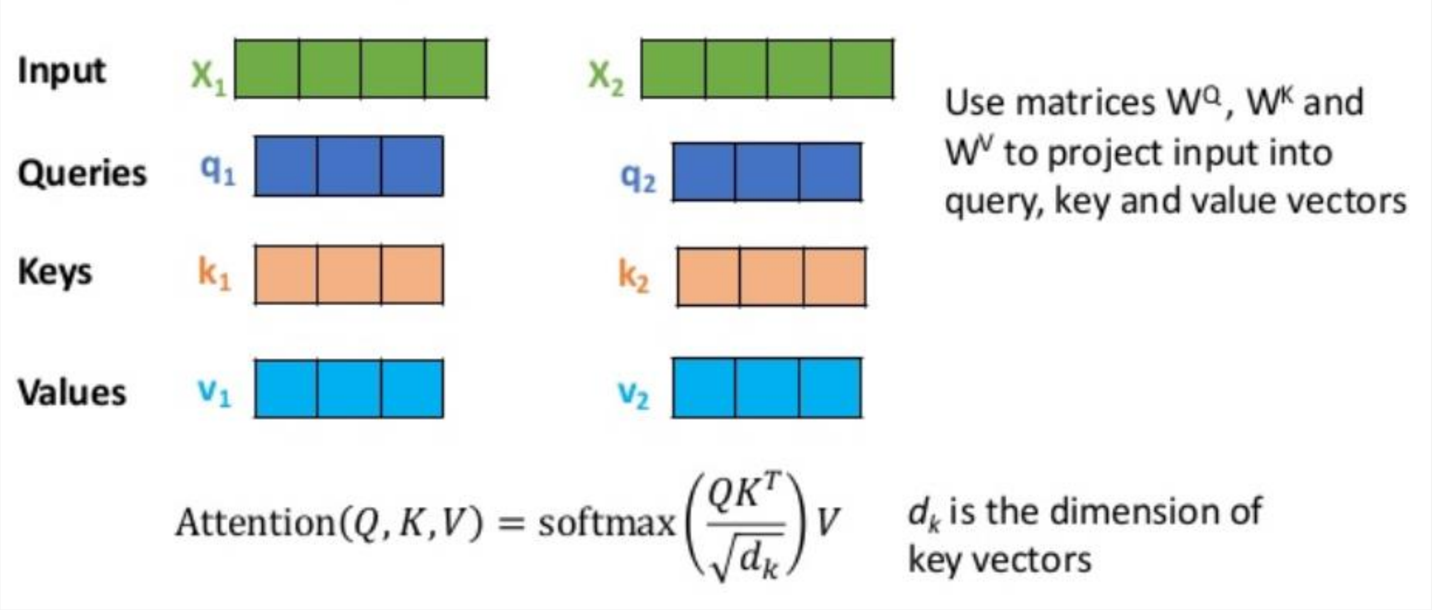
Name	Definition	Citation
Self Attention	Relating different positions of the same input sequence	Cheng2016
Global/Soft	Attending to the entire input state space	Xu2015
Local/Hard	Attending to the part of the input state space	Xu2015, Luong2015





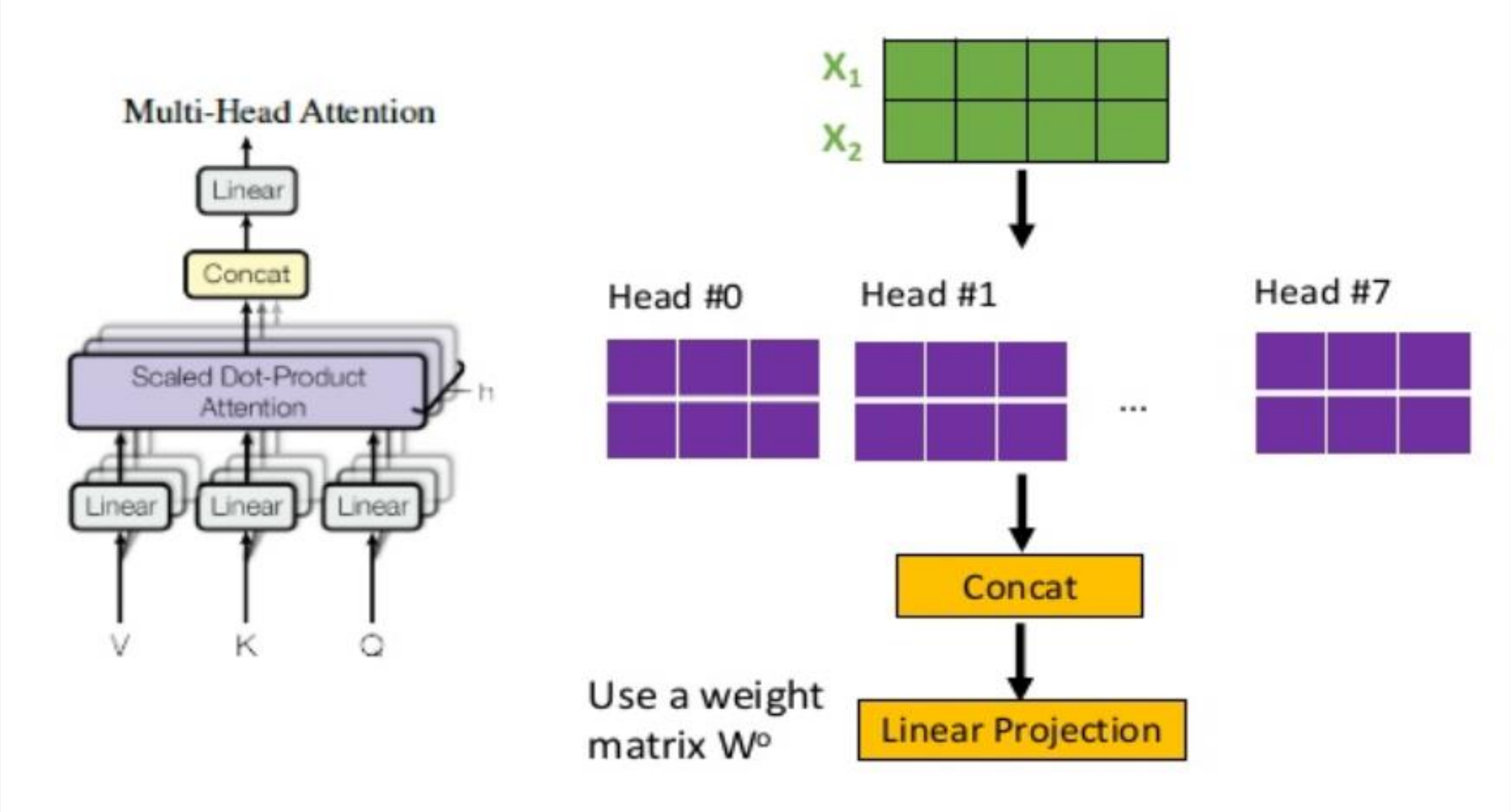
SELF-ATTENTION IN DETAIL

- Attention maps a query and a set of key-value pairs to an output
    - Query, keys and output are all vectors



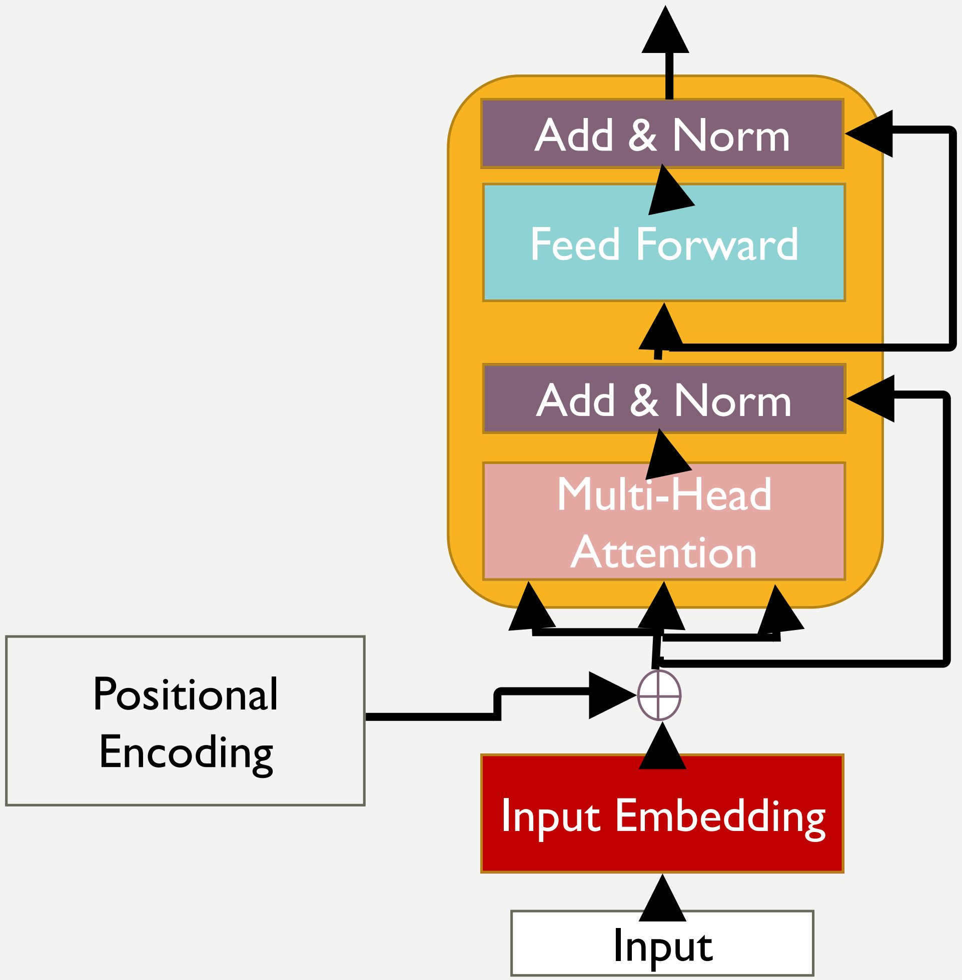
Q: Previous Decoder Hidden state
K: Encoder Hidden State
V: Encoder Hidden State

MULTI-HEAD ATTENTION

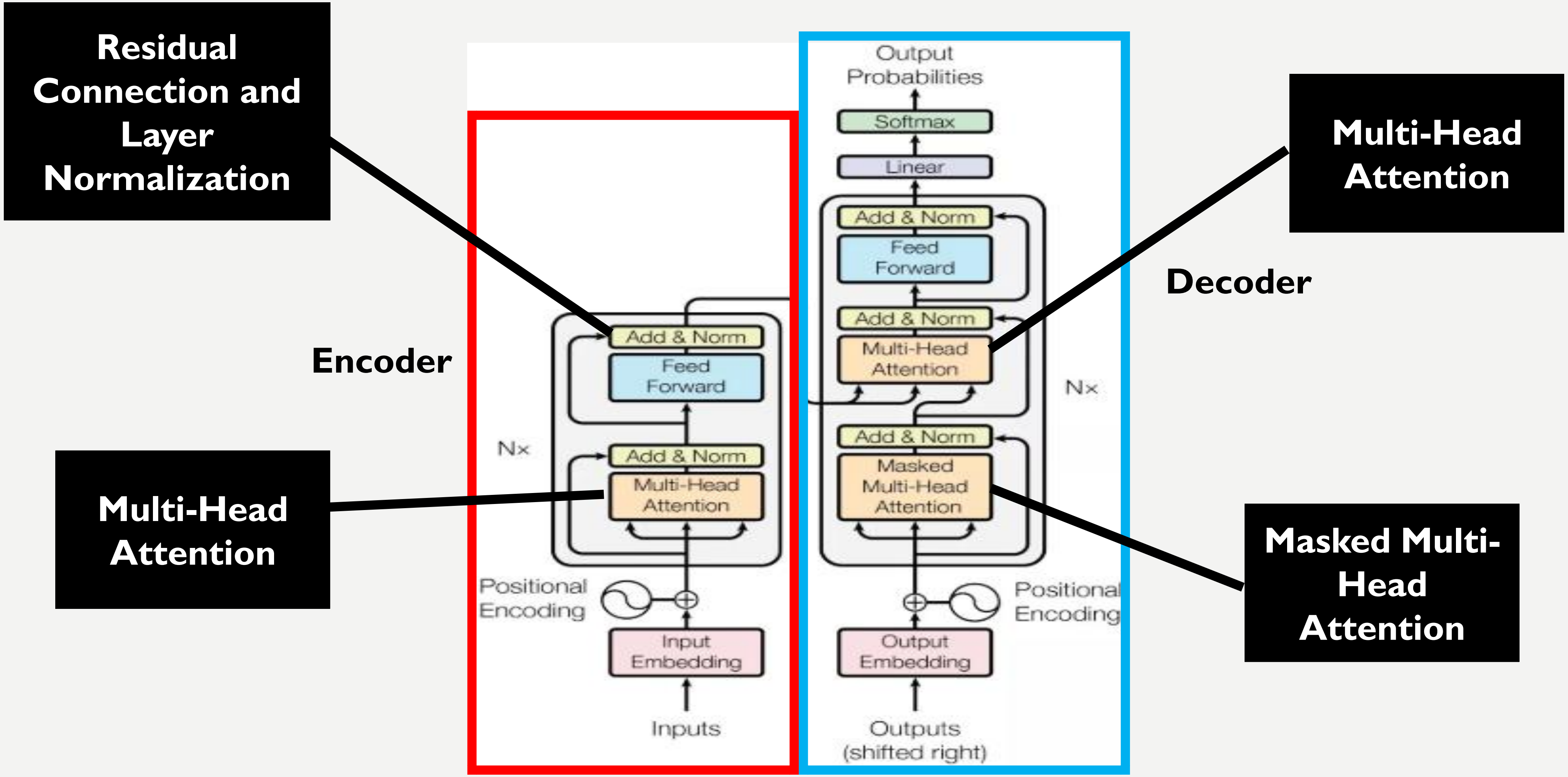


INSIDE AN ENCODER BLOCK

- In BERT experiments, the number of blocks N was chosen to be 12 and 24.
- Blocks do not share weights with each other



INSIDE AN ENCODER DECODER BLOCK



POSITION ENCODING

- Position Encoding is used to make use of the order of the sequence
  - Since the model contains no recurrence and no convolution
- In Vawasni et al., 2017, authors used sine and cosine functions of different frequencies

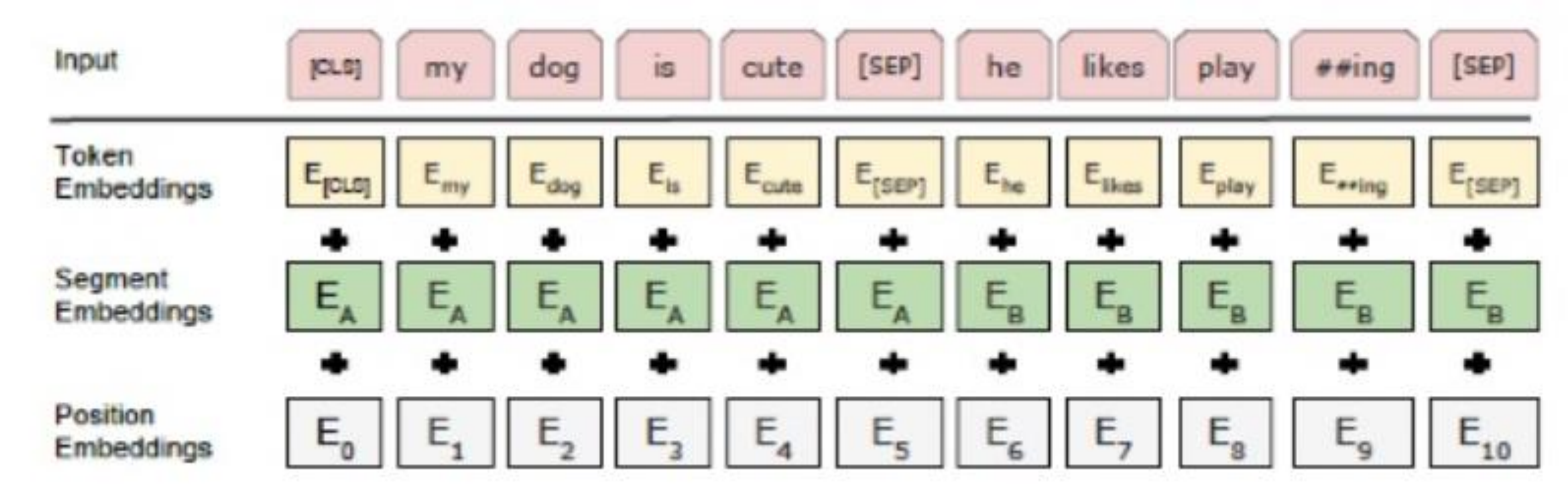
$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

- *pos* is the position and i is the dimension

INPUT REPRESENTATION

- Token Embedding: Use pre-trained WordPiece embedding.
- Position Embedding: Use learned Position Embedding
- Added sentence embedding to every tokens of each sentence.
- Use [CLS] for the classification tasks
- Separate sentences by using a special token [SEP]



PRE-TRAINING PROCEDURE

- Training data: BooksCorpus (800M words) + English Wikipedia (2,500M words)
- To generate each training input sequences: sample two spans of text (A and B) from the corpus
  - The combined length is < 500 tokens
  - 50% B is the actual next sentence that follows A and 50% of the time it is a random sentence from the corpus
- The training loss is the sum of the mean masked LM likelihood and the mean next sentence prediction likelihood

TASK#1: MASKED LM

- 15% of the words are masked at random
- Not all tokens were masked in the same way (example sentence "My dog is hairy")
  - 80% were replaced by the <MASK> token: "My dog is <MASK>"
  - 10% were replaced by a random token: "My dog is apple"
  - 10% were left intact: "My dog is hairy"

TASK#2: NEXT SENTENCE PREDICTION

- Motivation: Many downstream tasks are based on understanding the relationship between two text sentences
  - Question Answering (QA) and Natural Language Inference (NLI)
- Language modeling does not directly capture that relationship
- The task is pre-training binarized next sentence prediction task

Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]

Label = isNext

Input = [CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]

Label = NotNext

FINE-TUNING PROCEDURE

- **Classification tasks** such as sentiment analysis are done similarly to Next Sentence classification, by adding a classification layer on top of the Transformer output for the [CLS] token.
- In **Question Answering tasks** (e.g. SQuAD v1.1), the software receives a question regarding a text sequence and is required to mark the answer in the sequence. Using BERT, a Q&A model can be trained by learning two extra vectors that mark the beginning and the end of the answer.
- In **Named Entity Recognition** (NER), the software receives a text sequence and is required to mark the various types of entities (Person, Organization, Date, etc) that appear in the text. Using BERT, a NER model can be trained by feeding the output vector of each token into a classification layer that predicts the NER label.

COMPARISON OF BERT AND OPENAI GPT

OpenAI GPT	BERT
Trained on BookCorpus (800M)	Trained on BooksCorpus (800M) + Wikipedia (2,500M)
Use sentence separator ([SEP]) and classifier token ([CLS]) only at fine-tuning time	BERT learns [SEP], [CLS] and sentence A/B embedding during pre-training
Trained for 1M steps with a batch-size of 32,000 words	Trained for 1M steps with a batch-size of 128,000 words
Use the same learning rate of 5e-5 for all fine-tuning experiments	BERT choose a task-specific learning rate which performs the best on the development set

RESULTS

GLUE RESULT

Question Natural Language Inference is a
version of Stanford Question Answering Dataset
which
classificat
pairs whic
negative ex
whi

Multi-Genre Natu
classificatio
goal is to pre
an entailm

cl
pre
a

Star
is
class
se

Corpus of Linguistic Acceptability
is a
classifi
is to p
s

Seman
Benchm
senten
news

N
Parap
autom
o

Recognizing Textual
Entailment is a binary
entailment task similar to
MNLI but with less training
data

- GLUE benchm

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\mathrm{BERT}_{\mathrm{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\mathrm{BERT}_{\mathrm{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

SQUAD V1.1

- Stanford Question Answering Dataset is a collection of 100k crowdsourced question/answer pairs.
- Given a question and a passage from Wikipedia containing the answer, the task is to predict the answer text span in the passage.

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

CONCLUSIONS

- Unsupervised pre-training (pre-training language model) is increasingly adopted in many NLP tasks
- Major contribution of the paper is proposed a deep bidirectional architecture from Transformer
  - Advance state-of-the-art for many important NLP tasks

REFERENCES

- Jacob Devlin , Google AI Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"
- Ashish Vaswani et al., Cornel University, "Attention is all you need", 2017.
- https://zhuanlan.zhihu.com/p/47812375
- https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

谢谢！

ANY QUESTION??